CONFIDENCE INTERVAL

(Part 1)

Known standard deviation of population

RANDOM SAMPLING

Conditions to be met:

a. Every case in the population must have an *equal chance* of being selected

b. The selection of a case can in no way *affect* the selection of any other case

    Ex: Alternately selecting a Republican and Democrat violates this condition

c. Cases must be selected in such a way that *all combinations are possible*

Example: 100 marbles in a box; **50 red**, **50 blue**

Blindfolded, you draw 10 marbles randomly

1. Each has equal chance
2. Selecting red does not affect next marble
3. Any combination is possible

0 0 0 0 0 0 0 0 0 0 → **5R 5B**

0 0 0 0 0 0 0 0 0 0 → **6R 4B**

0 0 0 0 0 0 0 0 0 0 → **3R 7B**

0 0 0 0 0 0 0 0 0 0 → **4R 6B**

SAMPLING ERROR

Population may be large

Sample is representative of population

What is the average age of university students taking six hours or more of coursework?

Population = 10,000 students

Sample = 200 students

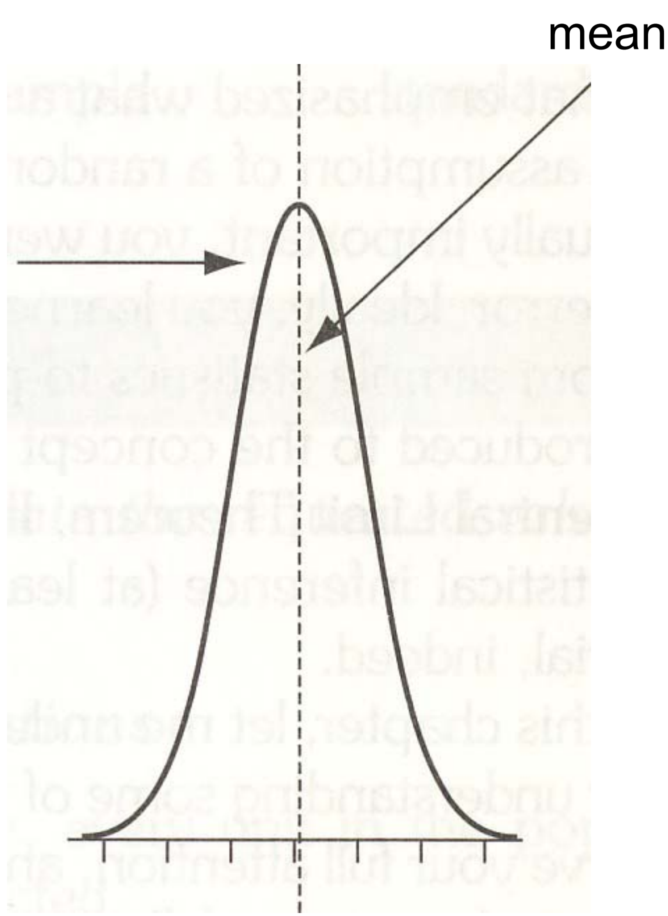
Ages range from very young to very old

(Infinite number of sample outcomes)

Mean of sample may not equal mean of population

SAMPLING DISTRIBUTION OF SAMPLE MEANS

Run sample 1000x and calculate/plot means



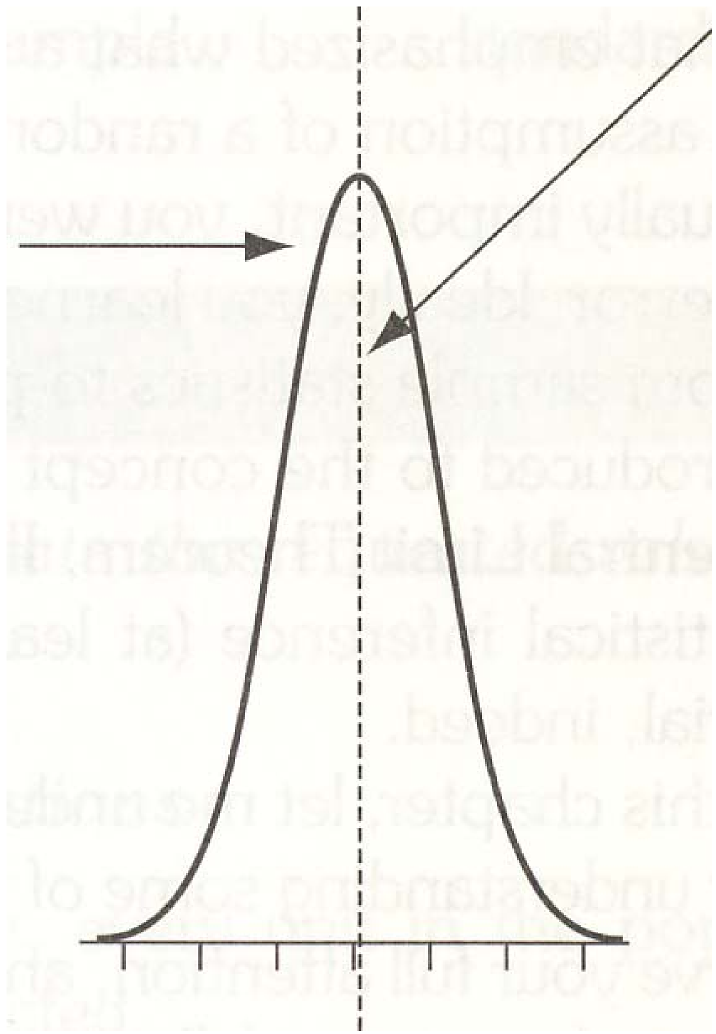
calculate mean of sample means

standard deviation of sample means

= ***standard error of the mean***

CENTRAL LIMIT THEOREM

"If repeated random samples of size ***n*** are taken from a population with a mean (μ) and a standard deviation (σ), the sampling distribution of sample means will have a mean equal to μ and a standard error equal to σ/sqrt(***n***). Moreover, as ***n*** increases, the sampling distribution will approach a normal distribution."


Mean
Many samples

What is the average age of university students taking six hours or more of coursework?

Cannot assume the sample mean = population mean

however

Might say, "*I believe the mean age (μ) of the university is between 23.4 and 26.1 years*." based on our sample.

Confidence Interval for the Mean

Confidence Interval for the Mean of population with known σ

SAT Scores --- $\mu = 500$, $\sigma = 100$

Assume a sample:

$n = 225$ $x = 606$ ($N = 10{,}000$ $\mu = ?$)

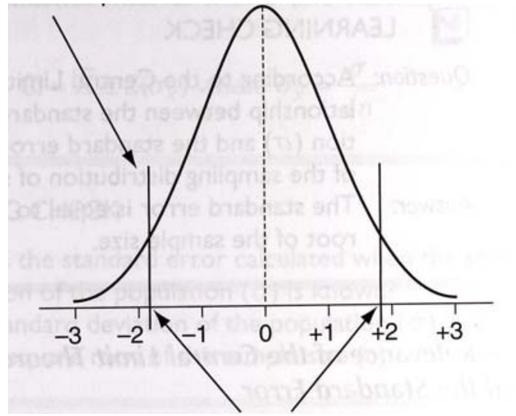
C.I. = Sample Mean ± Z ×(?)

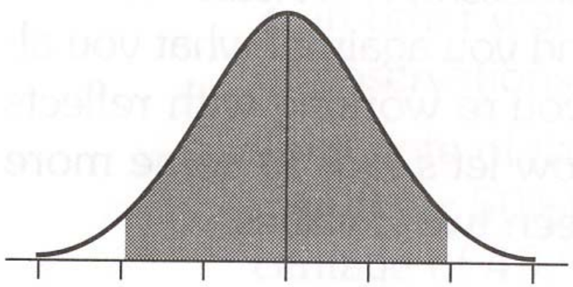
where Z = 1.96 for 95% confidence interval, or

2.58 for 99% confidence interval

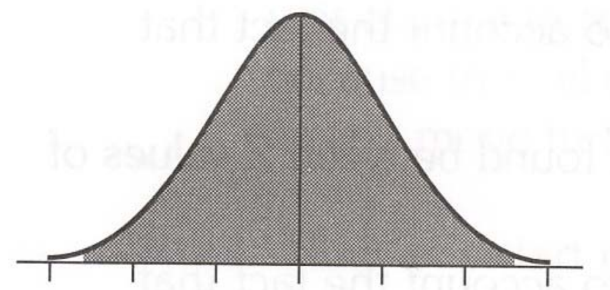
and (?) is the Standard Error of the Mean

Review of Z score





Z=1.96 (95%)



Z=2.58 (99%)

C.I. = Sample Mean $\pm$ Z $\times$(?)

(?) = $\sigma/\sqrt{n}$

$\sigma = 100$ $n = 225$ $x = 606$ $Z = 2.58$ (for 99%)

(?) = $100/\sqrt{225} = 100/15 = 6.67$

C.I. = 606 ± (2.58 x 6.67)

C.I. = 606 ± 17.21

C.I. = 588.79 to 623.21

Translation: *<u>99</u> times out of <u>100</u> our results would contain the mean of the population*

(588.79 – 623.21)

What if we set our confidence level at 95%?

$$CI = 606 \pm 1.96 \times (100/\sqrt{225})$$

$$= 606 \pm 1.96 \times 6.67$$

$$= (592.93 \text{ to } 619.07)$$

hence, *<u>95</u> times out of <u>100</u> our results would contain the mean of the population*

CONFIDENCE INTERVAL

(Part 2 of 2)

Unknown standard deviation of population

Confidence Interval for the Mean

of population with known σ

SAT Scores --- $\sigma = 100$

Assume a sample:

$n = 225 \quad x = 606 \qquad (N = 10{,}000 \quad \mu = ?)$

C.I. = Sample Mean ± Z ×(?)

where Z = 1.96 for 95% confidence interval, or

2.58 for 99% confidence interval

and (?) is the Standard Error of the Mean

C.I. = Sample Mean $\pm$ Z $\times$ $(\sigma/\sqrt{n})$

$\sigma = 100$ $n = 225$ $x = 606$ $Z = 2.58$ (for 99%)

C.I. = $606 \pm (2.58 \times 100/\sqrt{225})$

C.I. = 606 ± 17.21

C.I. = 588.79 to 623.21

Translation: *99 times out of 100 our results would contain the mean of the population somewhere between* 588.79 – 623.21

Confidence Interval with σ Unknown

Two problems:

1. Unknown σ
2. Cannot rely on Normal Curve

Use ***estimate of the standard error of the mean***

s_x = sample std dev / square root of n

$$s_x = s / \sqrt{n}$$

Example: *Want to know the average expenditure per customer in the bookstore*

Sample size = 100

Mean = $31.50

Std Dev = $4.75

$s_x = s / \sqrt{n} \quad = 4.75 / \sqrt{100} \quad = 4.75 / 10 = .475$

$s_x = 0.48$ (Est of std error of mean)

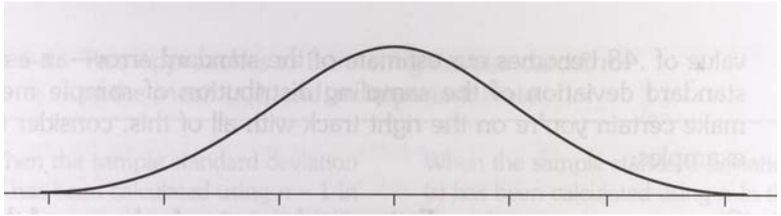
2nd problem: Don't know population mean or std dev, thus cannot use normal distribution or Z value

Not to fear; statisticians to the rescue:

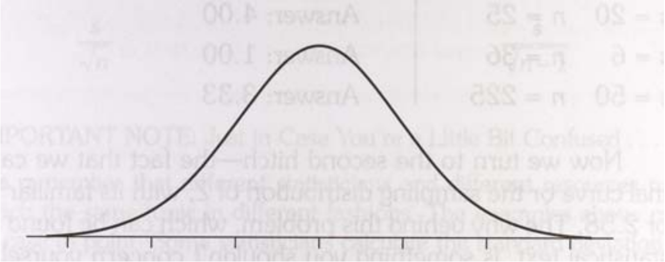
family of *t* distributions

Gossett who worked for Guiness Brewery developed the concept

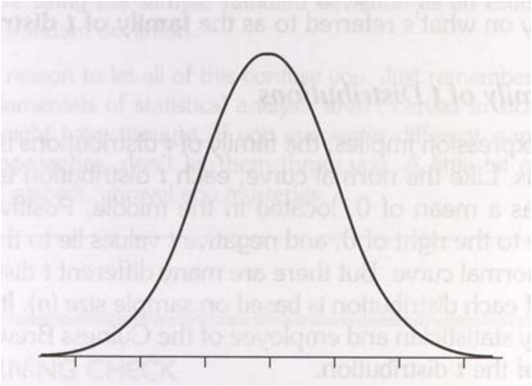
***t* distribution** :"the shape of a sampling distribution depends on the number of cases in each of the cases"



Small sample size



Larger sample size



Still larger sample size

Similar to Z values

Another concept: **Degrees of Freedom**

In a distribution of *n* cases, *n*-1 cases are free to vary

Example: Quiz worth 10 points

Five scores (*n*=5), Mean = 8,

Four (*n*-1) of the five are free to vary

Assume four scores are 8, 8, 10, 10

(They could have been anything from 0-10)

Once the four are known, the fifth one cannot vary

(It must be 4)

(5x8=40) & (8+8+10+10=36) therefore 40-36 = 4

Degrees of Freedom

(n-1)

Degrees of Freedom	LEVEL OF SIGNIFICANCE .20	.10	.05	.02	.01	.001
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
60	1.296	1.671	2.000	2.390	2.660	3.460
80	1.292	1.664	1.990	2.374	2.639	3.416
100	1.290	1.660	1.984	2.364	2.626	3.390
120	1.289	1.658	1.980	2.358	2.617	3.373
Infinity	1.282	1.645	1.960	2.327	2.576	3.291

Level of Significance

95% Confidence Interval (1-.05)

Degrees of Freedom	LEVEL OF SIGNIFICANCE .20	.10	.05	.02	.01	.001
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
60	1.296	1.671	2.000	2.390	2.660	3.460
80	1.292	1.664	1.990	2.374	2.639	3.416
100	1.290	1.660	1.984	2.364	2.626	3.390
120	1.289	1.658	1.980	2.358	2.617	3.373
Infinity	1.282	1.645	1.960	2.327	2.576	3.291

Degrees of Freedom	LEVEL OF SIGNIFICANCE					
	.20	.10	.05	.02	.01	.001
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
60	1.296	1.671	2.000	2.390	2.660	3.460
80	1.292	1.664	1.990	2.374	2.639	3.416
100	1.290	1.660	1.984	2.364	2.626	3.390
120	1.289	1.658	1.980	2.358	2.617	3.373
Infinity	1.282	1.645	1.960	2.327	2.576	3.291

99% Confidence Interval

(1 - .01)

EXERCISE:

Random sample of 25 retirees. Want to estimate the average number of emails sent out each week. Our sample provides a mean of 12 (emails per week) with a standard deviation of 3. We decide to construct a 95% confidence interval.

$$\textbf{CI = Sample Mean} \pm t(s_x)$$

(Remember: $s_x = s / \sqrt{n}$)

$$\textbf{CI} = 12 \pm t(3/\sqrt{25})$$

$n = 25$

$df = 24$ (n-1)

95% Confidence Interval = .05 Level of Significance

$t = 2.064$

Degrees of Freedom (df)	LEVEL OF SIGNIFICANCE					
	.20	.10	.05	.02	.01	.001
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
50	1.299	1.676	2.009	2.403	2.678	3.496
60	1.296	1.671	2.000	2.390	2.660	3.460
80	1.292	1.664	1.990	2.374	2.639	3.416
100	1.290	1.660	1.984	2.364	2.626	3.390
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.327	2.576	3.291

CI = Sample Mean ± *t*(s / $\sqrt{n}$)

CI = 12 ± 2.06(3/ $\sqrt{25}$)

CI = 12 ± 2.06(3/5)

CI = 12 ± 2.06(.6)

CI = 12 ± 1.24

CI = 10.76 *to* **13.24**

Our population mean is between **10.76** & **13.24**. And, our method will produce a <u>correct estimate 95 out of 100 times</u>.

n = 25

df = 24 (*n*-1)

99% Confidence Interval = .01 Level of Significance

***t* = 2.797 (or 2.80)**

Degrees of Freedom (*df*)	LEVEL OF SIGNIFICANCE					
	.20	.10	.05	.02	.01	.0
5	1.476	2.015	2.571	3.365	4.032	6.
6	1.440	1.943	2.447	3.143	3.707	5.
7	1.415	1.895	2.365	2.998	3.499	5.
8	1.397	1.860	2.306	2.896	3.355	5.
9	1.383	1.833	2.262	2.821	3.250	4.
10	1.372	1.812	2.228	2.764	3.169	4.
11	1.363	1.796	2.201	2.718	3.106	4.
12	1.356	1.782	2.179	2.681	3.055	4.
13	1.350	1.771	2.160	2.650	3.012	4.
14	1.345	1.761	2.145	2.624	2.977	4.
15	1.341	1.753	2.131	2.602	2.947	4.
16	1.337	1.746	2.120	2.583	2.921	4.
17	1.333	1.740	2.110	2.567	2.898	3.
18	1.330	1.734	2.101	2.552	2.878	3.
19	1.328	1.729	2.093	2.539	2.861	3.
20	1.325	1.725	2.086	2.528	2.845	3.
21	1.323	1.721	2.080	2.518	2.831	3.
22	1.321	1.717	2.074	2.508	2.819	3.
23	1.319	1.714	2.069	2.500	2.807	3.
24	1.318	1.711	2.064	2.492	2.797	3.
25	1.316	1.708	2.060	2.485	2.787	3.
26	1.315	1.706	2.056	2.479	2.779	3.
27	1.314	1.703	2.052	2.473	2.771	3.
28	1.313	1.701	2.048	2.467	2.763	3.
29	1.311	1.699	2.045	2.462	2.756	3.
30	1.310	1.697	2.042	2.457	2.750	3.
40	1.303	1.684	2.021	2.423	2.704	3.5
50	1.299	1.676	2.009	2.403	2.678	3.4
60	1.296	1.671	2.000	2.390	2.660	3.4
80	1.292	1.664	1.990	2.374	2.639	3.4
100	1.290	1.660	1.984	2.364	2.626	3.3
120	1.289	1.658	1.980	2.358	2.617	3.3
∞	1.282	1.645	1.960	2.327	2.576	3.2

CI = Sample Mean ± *t*(s / $\sqrt{n}$)

CI = 12 ± 2.80(3/ $\sqrt{25}$)
CI = 12 ± 2.80(3/5)
CI = 12 ± 2.80(.6)

CI = 12 ± 1.68

CI = 10.32 *to* **13.68**

Our population mean is between **10.32** & **13.68**. And, our method will produce a <u>correct estimate 99 out of 100 times</u>.

Final note:

A confidence interval for the mean ***<u>does not</u>*** provide you with an exact estimate of the population mean.

Rather, it provides you with an interval that you believe contains the true mean,

and you are confident that 95% (*or 99%, etc*) of your confidence intervals would contain the true mean.